Available online at www.sciencedirect.com
ScienceDirect
Theoretical Computer Science 374 (2007) 111–126
Theoretical Computer Science
www.elsevier.com/locate/tcs

Reversal and transposition medians

Niklas Eriksen*

Mathematical Sciences, Göteborg University and Chalmers University of Technology, SE-412 96 Göteborg, Sweden

Received 28 June 2005; received in revised form 3 May 2006; accepted 12 December 2006

Communicated by M. Crochemore

Abstract

In determining phylogenetic trees using gene order information, medians provide a powerful alternative to pairwise distances. On the other hand, both breakpoint and reversal medians are NP-hard to compute and the use of medians has been limited to relatively closely related genomes. In this paper, we show that in spite of the greater non-uniqueness of reversal medians, compared to breakpoint medians, medians of moderately distant genomes are often widely spread. This means that regardless of which approximation algorithms one may devise for computing reversal medians, the genomes need to be closely related for phylogenetic tree computations to be successful. To show this, we use results on transposition medians, which behave similarly, and also support our claims with simulations and a real data example with widely spread medians.
© 2006 Elsevier B.V. All rights reserved.

Keywords: Median; Reversal; Transposition; Genome rearrangement

1. Introduction

Comparative genomics strives at describing how different species are related to each other, and how they have evolved. Optimally, the resulting object is a phylogenetic tree, which is a family tree for species. During the last decades, many algorithms and mathematical tools have been invented that use genomic data to say something about the phylogenetic tree of a set of species.

Concentrating on gene order data, much effort has been put on computing the evolutionary distance between species. There are several ways to estimate the distance (for explanations of unknown terms, see Section 2), for instance counting breakpoints (easy) [24], using the reversal distance (linear time) [1,2,4,18], the expected reversal distance (polynomial time) [13,15,29], the (block) transposition distance (unknown complexity) [3,10,12] and combinations of several operations [6,14]. If the evolutionary distances between all species in a set have been computed correctly, the correct tree can be easily deduced.

Another approach is using medians. This should in general give stronger results, giving information not only about the topology and edge lengths of the tree, but also the gene order of the ancestral species. However, while distance computations tend to be polynomially computable, or at least given to good approximations, medians are in general NP-hard to compute. This is true not only for the reversal median [8], but also for the breakpoint median [23].

* Tel.: +46 31 772 35 99.
E-mail address: ner@math.chalmers.se.

0304-3975/$ - see front matter © 2006 Elsevier B.V. All rights reserved.
doi:10.1016/j.tcs.2006.12.006

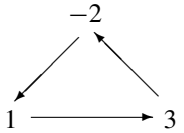


Fig. 1. The genome [1 3 −2].

Still, we have seen several attempts to use medians in tree building. Blanchette and Sankoff [24] reduced the BREAKPOINT MEDIAN PROBLEM to the TRAVELLING SALESPERSON'S PROBLEM, for which several practical approximation algorithms exist. With these, they proposed an algorithm that for each possible tree topology used median computations to compute the tree ancestors. Then, the tree with the lowest sum of edge lengths was chosen as the best. Their implementation was improved on by Moret et al. [19,21,22], who also added the option of using reversal medians. While significantly slower to compute than breakpoint medians, the reversal median algorithms of Caprara [9] and Siepel [26,27] gave much more relevant medians, thus improving on the total computational time.

Another, iterative approach to phylogeny reconstruction has been proposed by Bourque and Pevzner [7] and Eriksen [16, Ch. 9]. Given a tree, adding a new genome amounts to splitting one of the edges in the tree and adding a median there. Median computations are based on the idea that performing reversals on one genome that decrease the distance to the other genomes should also decrease the distance to a median. This is a good idea, but it relies on our ability to choose the best reversal that fulfills this condition, to maximise the number of them we can perform.

In general, reversal median algorithms tend to work well only for closely related species. In particular, the algorithm of Siepel shows really bad running time behaviour as the distance between the species increases [20]. If the distance between the species is reasonably large, it seems that these algorithms will not be of much help to us. We ask ourselves if this is inherent to the problem or if these algorithms can be improved.

In this paper, we show that when the distance between the species is reasonably large, there will never be a useful reversal median algorithm. The reason is that the reversal median is far from unique, even for moderately large distances. For instance, we show with simulation on a genome with $n = 40$ genes that if the evolutionary distance from the ancestor to three given genomes is 15 reversals each, the reversal medians of these genomes may be as far apart as 20 reversals, none of them equal to the ancestor. Similarly, we give a real data example where two candidate medians are twice as far apart as their distance to the closest neighbour in the test set. Not knowing which of these medians best approximates the ancestor and just picking one may introduce large propagating errors into the tree.

We achieve our goal by deriving results for ordinary permutations, using (ordinary) transpositions as operations. By an argument given in [15], this model shows very similar behaviour to the reversal model, especially concerning expected distances after $t < n$ random operations. Thus, the results we obtain for the symmetric group carry over to the genome setting. In particular, we show that while the distance between the ancestor and the descendants in general has increased by each applied operation for medium distances, this could not be said about the distance between the respective descendants. As a consequence, the medians will be widely spread and far from the ancestor.

2. A background to genome rearrangement problems

An **evolutionary tree** is a tree in the graph-theoretical sense, with present species as leaves and extinct ancestors as inner nodes. The tree could be rooted, with a time line, but we focus solely on undirected trees in this article. To each edge is associated an evolutionary distance. In our context, we estimate this distance using gene order information.

A **genome with n genes** is a signed, circular, permutation on n elements, that is an ordinary permutation $\pi \in \mathfrak{S}_n$ which has only one cycle, and where each element has a positive or negative sign attached to it. The set of all genomes with n genes is denoted G_n. For simplicity, the genome $\pi \in G_n$ is often written in a linear fashion: $\pi = [\pi_1\ \pi_2\ \ldots\ \pi_n]$. It is then understood that the leftmost gene should be attached to the rightmost gene. The identity genome is denoted $\mathrm{id} = [1\ 2 \ldots n]$.

Example 1. The genome in Fig. 1 can be written as, for instance, [1 3 −2] or [3 −2 1] or even [−3 −1 2] (reading in the opposite direction). Usually, we let 1 be the first element in the linear order.

There are several evolutionary events that may change a genome, two of which are depicted in Fig. 2. A **reversal** (or sometimes **inversion**) between π_i and π_j, where $i \neq j$, is an operation that takes the segment $\pi_{i+1}\pi_{i+2} \ldots \pi_j$ out of the genome and inserts it at the same place backwards, changing the sign of all elements in the segment. A **block transposition** between π_i, π_j and π_k, where $i \neq j \neq k \neq i$, is an operation that takes the segment $\pi_{i+1}\pi_{i+2} \ldots \pi_j$ out of the

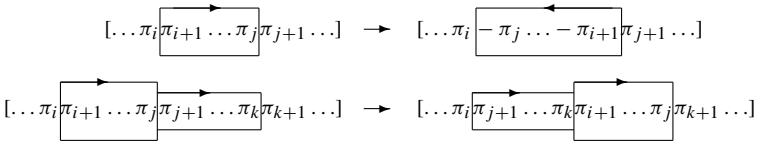


Fig. 2. Definitions of reversal and block transposition on genomes.

genome and inserts it directly after π_k. The block transposition often goes by the name transposition, but we will use the full name here to separate it from the transposition on the symmetric group, which is frequently mentioned below.

Given any distance $d(a, b)$ on a set S, we define a **median** of a subset (the **test set**) $\{s_1, s_2, \ldots, s_k\} \subset S$ to be an $s \in S$ such that the **median distance** $\sum_i d(s, s_i)$ is minimised. The algorithmic problem MEDIAN is that of finding a median of a given set. Note that each distance induces a median problem in its own right. In this paper, we will study the problems TRANSPOSITION MEDIAN (for the symmetric group) and REVERSAL MEDIAN. Also note that MEDIAN is trivial for $k = 2$ with any distance, and in general hard for $k \geq 3$.

3. The analogy between reversal and transposition distances

In order to say something about the symmetric group $\mathfrak{S}_n$, one commonly considers its generators, the **adjacent transpositions** $(i\ (i+1))$ (in cycle notation), or their conjugates, the **transpositions** $(i\ j)$. Genomes may be seen as a subset of the signed permutation on n elements, but it is not a subgroup under composition, and there is no clear relation between reversals and block transpositions on the one hand, and (adjacent) transpositions on the other. Thus, we can not analyze G_n using our usual combinatorial and algebraic means.

However, there are major similarities between the behaviour of G_n and $\mathfrak{S}_n$ with respect to reversal and transposition distances. We shall derive results about $\mathfrak{S}_n$ using standard combinatorial techniques, and argue that these are also applicable for G_n. We will use $d_{\text{rev}}(\pi, \tau)$ to denote the reversal distance between genomes π and τ, and $d_{\text{trp}}(\pi, \tau)$ to denote the transposition distance between permutations π and τ. Since the transposition distance between π and τ equals the transposition distance between $\pi\tau^{-1}$ and id, we will often use the notation $d_{\text{trp}}(\pi) = d_{\text{trp}}(\pi, \text{id})$. Also, the set of sequences of exactly t transpositions in $\mathfrak{S}_n$ is denoted

$$\mathcal{P}_{nt} = \{(i_1\ j_1)(i_2\ j_2)\ldots(i_t\ j_t) : 1 \leq i_k < j_k \leq n, 1 \leq k \leq t\}.$$

The close relationship between G_n and $\mathfrak{S}_n$ is discussed in [15]. It is well-known that permutations have an interesting cycle structure, and there is a similar cycle structure associated to each genome. If $c(\pi)$ denotes the number of cycles in π, then we have $d_{\text{trp}}(\pi) = n - c(\pi)$ for $\pi \in \mathfrak{S}_n$ and $d_{\text{rev}}(\sigma, \text{id}) \approx n - c(\sigma)$ for $\sigma \in G_n$. But there is more to it than this superficial resemblance. In short, Eriksen and Hultman establish that while a transposition $(a\ b)$ will increase the distance (to id) if a and b belong to different cycles and decrease the distance if they belong to the same cycle, the reversal $a \ldots b$ will in general do the same. Exceptions may turn up if a and b belong to the same cycle. These exceptions are not uncommon if a random genome is picked, but for genomes fairly close to id, it is unlikely that a and b belong to the same cycle, and even then, chances are small that the distance is not reduced.

We find that if we apply a random transposition to a permutation π and the corresponding reversal to a genome σ with the same cycle structure (that is there exists a length preserving bijection between the cycles of π and the cycles of σ), then the distances to the identities will in most cases change by an equal amount. This approximation holds particularly well for permutations and genomes close to the identity. It seems reasonable that the expected distances after t operations will be approximately equal, at least for $t < n$, say. In fact, simulations carried out in [15] indicate that these are close up to at least $t = 3n/2$, after which the reversal distance gives very little information.

Example 2. We look at the transposition poset for $\mathfrak{S}_3$ and the reversal poset for G_3 in Fig. 3. In these posets, y covers x if and only if there is a transposition (reversal) transforming x to y and $d(x) < d(y)$, where $d(x)$ refers to the respective distances. Both posets are ranked by the distance to id. We see that at the two bottom levels, the probability to move up or down is the same, if a transposition or a reversal, respectively, is taken randomly from a uniform distribution. On the third level, where the distance is two, the probabilities are fairly similar, the difference being that the permutations have reached their maximal level, but the genomes have not.

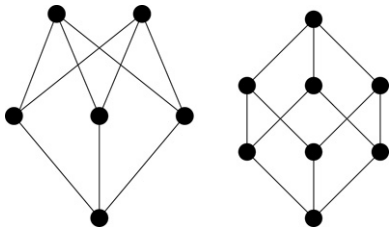


Fig. 3. The transposition poset of $\mathfrak{S}_3$ and the reversal poset for G_3. These are ranked posets, and the rank is the distance from id at the bottom. We observe that the probabilities to move up or down in these diagrams when applying random operations are the same for the first two levels. In general, these probabilities are very similar for most levels, the exceptions being positioned at the top.

While all elements in the transposition poset of $\mathfrak{S}_n$ has $\binom{n}{2}$ edges, $n \geq 1$, this is not true for the reversal poset of G_n. In fact, some reversals do not change the distance to id. For example, the genome $[1\ -5\ 4\ 2\ -3]$ has distance 4, as does $[1\ -4\ 5\ 2\ -3]$. Still, most reversals do change the distance of genomes, and the proportions of edges going up at some level are similar to the proportions for transpositions on $\mathfrak{S}_n$ at the same level.

4. Computing transposition medians

It is easy to compute fairly good upper and lower bounds for the transposition median distance. Given permutations π^1, π^2 and π^3, we can put one of these, say π^1, in the middle. This gives an upper bound on the median distance, namely $d(\pi^1, \pi^2) + d(\pi^1, \pi^3)$.

On the other hand, the triangle inequality must hold, so for any permutation π we have $d(\pi, \pi^1) + d(\pi, \pi^2) \geq d(\pi^1, \pi^2)$. In all, we get, for a median π,

$$\frac{\sum_{i<j} d(\pi^i, \pi^j)}{2} \leq \sum_i d(\pi, \pi^i) \leq \min_i \sum_j d(\pi^i, \pi^j).$$

This immediately gives a 4/3-approximation of the median distance, since

$$3 \min_i \sum_j d(\pi^i, \pi^j) \leq \sum_{i \neq j} d(\pi^i, \pi^j) = 4 \frac{\sum_{i<j} d(\pi^i, \pi^j)}{2}.$$

For k genomes, we get in a similar fashion,

$$\frac{\sum_{i<j} d(\pi^i, \pi^j)}{k-1} \leq \sum_i d(\pi, \pi^i) \leq \min_i \sum_j d(\pi^i, \pi^j),$$

for any median π. Again, this gives a fair approximation, in this case by the factor $2 - 2/k$. This shows that TRANSPOSITION MEDIAN is trivial for $k = 2$, as expected.

Taking the midmost of the given permutations thus gives a 2-approximation of the median. A naïve algorithm for computing a median of a set $S = \{\pi^1, \pi^2, \ldots, \pi^k\} \subset \mathfrak{S}_n$ would be to step by step improve optimally on this approximation, that is to apply transpositions to π^1, say, that reduce the distance to all other members of S. This is the approach taken in MGR by Bourque and Pevzner [7] and in Yggdrasil by Eriksen [16, Ch, 9]. Without loss of generality, we assume $\pi^1 = \mathrm{id}$. Returning to $k = 3$, the maximal number of transpositions that can be applied is

$$\frac{d(\pi^2) + d(\pi^3) - d(\pi^2, \pi^3)}{2}.$$

If we can find this number of transpositions that step by step decrease the distance to both π^2 and π^3, then we have found a median. Finding fewer steps, we can not be sure.

Example 3. Consider the three permutations $\pi^1 = \texttt{48715236}$, $\pi^2 = \texttt{43752861}$ and $\pi^3 = \texttt{37168524}$. We find that $d(\pi^1, \pi^2) = 5$, $d(\pi^1, \pi^3) = 7$ and $d(\pi^2, \pi^3) = 6$. This means that if we can apply

$$\frac{d(\pi^2, \pi^1) + d(\pi^2, \pi^3) - d(\pi^1, \pi^3)}{2} = \frac{5 + 6 - 7}{2} = 2$$

transpositions to π^2 that approach the other two permutations, then we have found a median. From the other permutations, we need to apply three and four transpositions, respectively.

Multiplying by the inverse of π^2 from the left, the other two permutations become $\texttt{16384527} = (265487)$ and $\texttt{23876451} = (1238)(4756)$. If we apply, for instance, the transpositions (5 6) and (4 7), this reduces the distance to both of these permutations by two, and thus we have a median. We could also have applied (5 6) and (2 8), which shows that the median is not unique.

On the other hand, consider the genomes $\pi^1 = \text{id}$, $\pi^2 = \texttt{214365} = (12)(34)(56)$ and $\pi^3 = \texttt{365214} = (135)(642)$. We would need to apply

$$\frac{d(\pi^1, \pi^2) + d(\pi^1, \pi^3) - d(\pi^2, \pi^3)}{2} = \frac{3 + 4 - 3}{2} = 2$$

transpositions to be sure to have a median. But we can not even apply one! Actually, the median distance equals the upper limit given above, and π^2 is a median.

Still, for $k = 3$ and moderately distant permutations, relative the number of genes, finding the maximal number of transpositions that decrease the distance of id to both other permutations, or equivalently transpositions that decrease the distance to id for both other permutations, usually gives the median. We name this problem of finding the longest common sequence of cycle splitting transpositions in a set $\{\pi_2, \ldots, \pi_k\}$ of permutations MAXIMAL APPROACH. Unfortunately, this problem is hard, at least when the total number of permutations, including id, is greater than 3.

Theorem 4. *The problem* MAXIMAL APPROACH *is NP-complete for* $k \geq 4$.

Proof. We show this by reducing MAXIMAL INDEPENDENT SET to MAXIMAL APPROACH. Assume that we have an undirected graph $G = (V, E)$. To each vertex $v_i \in V$, we associate the transposition $(a_i\ b_i)$. We shall now construct

$$k = \max_{v \in V} \left\lceil \frac{d(v)}{2} \right\rceil + 2$$

permutations, where $d(v)$ is the degree of v, and show how solving MAXIMAL APPROACH will give the solution to MAXIMAL INDEPENDENT SET.

Put $\pi_1 = \text{id}$ and $\pi_2 = (a_1\ b_1)(a_2\ b_2) \ldots (a_{|V|}\ b_{|V|})$. For the remaining permutations, proceed as follows. Divide the edges of G into $k - 2$ disjoint sets E_i such that $\bigcup E_j = E$ and $G_j = (V, E_j)$ has maximal degree 2. This can be done greedily and hence in linear time in $|E|$ and quadratic time in $|V|$.

The edges of E_i now form cycles and paths. For a cycle $v_1 v_2 \ldots v_m v_1$, we form the permutation cycle

$$(a_1\ b_m\ a_2\ b_1\ a_3\ b_2 \ldots a_{m-1}\ b_{m-2}\ a_m\ b_{m-1}),$$

and for the path $v_1 v_2 \ldots v_3$, we form the permutation cycle

$$(a_1\ a_2\ b_1\ a_3\ b_2 \ldots a_{m-1}\ b_{m-2}\ a_m\ b_{m-1}\ b_m).$$

All permutation cycles coming from E_j are multiplied to give π_{j+2}.

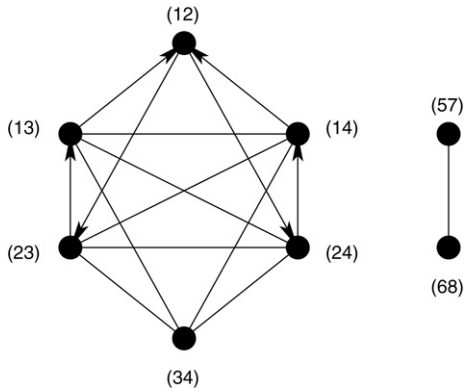
Any sequence of allowed transpositions on π_1 now corresponds to an independent subset of G, and a maximal approaching sequence corresponds to a maximal independent set. We have thus reduced MAXIMAL INDEPENDENT SET to MAXIMAL APPROACH. For graphs with maximal degree $\Delta \geq 3$, MAXIMAL INDEPENDENT SET is NP-complete [5]. Thus, MAXIMAL APPROACH is NP-hard for $k \geq 4$. Trivially, MAXIMAL APPROACH is in NP, and thus NP-complete, for $k \geq 4$.

We have not found a polynomial time algorithm for solving MAXIMAL APPROACH for $k = 3$. This is a problem shared with previous attempts, where the set of operations is reversals (MGR) and weighted combinations of reversals and block transpositions (Yggdrasil). In both these cases, this is partly dealt with by allowing suboptimal operations, as long as these lead to better operations afterwards. For transpositions, however, we can also use the fact that performing a valid transposition, that is a transposition that moves a genome closer to all other genomes, will never make

another transposition valid, as can be the case for reversals. We can therefore seek the longest sequence of valid transpositions for a genome by making a reduction to MAXIMAL ACYCLIC SUBGRAPH and apply a greedy algorithm for the latter problem.

MAXIMAL ACYCLIC SUBGRAPH is the problem of finding, in a directed graph, the maximal induced subgraph that does not contain any cycles. The graph $G = (V, E)$ is obtained as follows. Given a set of permutations $S_{\text{red}} = \{\pi_2, \ldots, \pi_k\}$, we let V be all pairs $(a\ b)$ of elements in $[n]$ such that $a \neq b$ and a and b are in the same cycle in all permutations in S_{red}. An edge is drawn from $(a_1\ b_1)$ to $(a_2\ b_2)$ if there is a permutation $\pi \in S_{\text{red}}$ such that applying $(a_1\ b_1)$ puts a_2 and b_2 in different cycles in π. Given any acyclic subgraph of G, we can always order the vertices such that no vertex has an edge directed at a vertex appearing later in that order. Consequently, all corresponding transpositions will be cycle splitting.

Example 5. Let us take a look at the example $\pi^2 = (1234)(5678)$ and $\pi^3 = (1243)(57)(68)$. There are eight transpositions that reduce the distance to id in both of these. They inhibit each other according to this graph, where doubly directed arrows are drawn without heads:

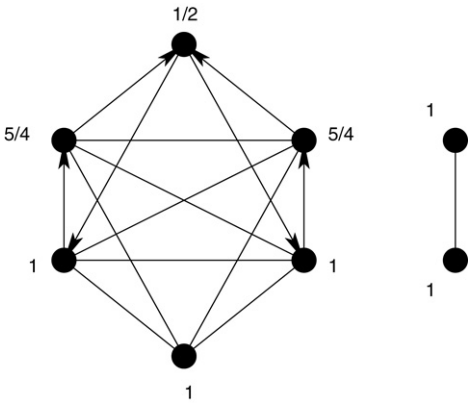


Note that while the outdegree equals the indegree for every vertex, we do get a directed graph. By inspection, we find a maximal acyclic subgraph to be generated by $V_{\text{MAS}} = \{(1\ 3), (1\ 2), (5\ 7)\}$.

To find an acyclic subset, we iterate over all vertices in V. Each vertex v gets the weight 1 and distributes this weight in equal parts to all vertices that point to v. This will give a large weight to those transpositions that inhibit many other transpositions, and a low weight to those that inhibit few, specially if those inhibit many other permutations. Then, we choose the vertex/transposition with the lowest weight, remove it along with the vertices it points at, and iterate.

This algorithm often gives the optimal solution for not too distant permutations, obtaining the lower limit for the median distance. If this limit is not obtained, it is hard to determine whether this is because the median distance is greater than the lower limit, or if the algorithm performs poorly. On occasions, we have seen an improvement by choosing a randomly chosen vertex of low weight, allowing weights up to 0.2 higher than the minimal. In most cases, however, this gives the same or slightly worse result.

Example 6. We continue the previous example. We show the same graph, but now with weights on the vertices:



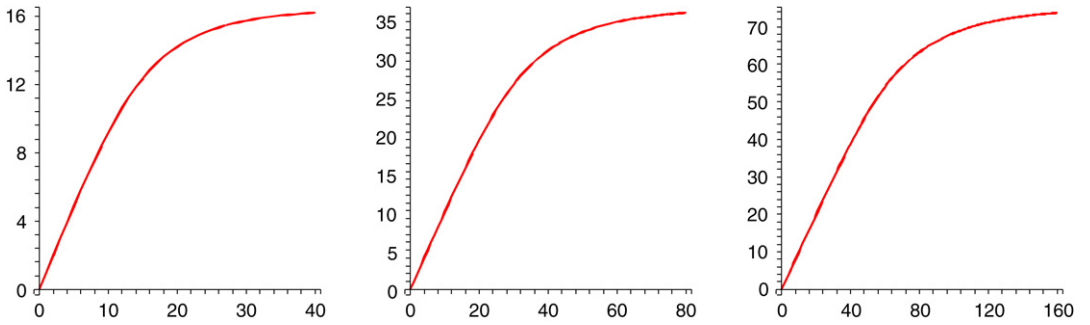


Fig. 4. The expectation of the transposition distance after t transpositions, $\mathbb{E}(X_t)$, with t on the abscissa. From left to right, we have 20 genes, 40 genes and 80 genes. The graphs are very similar, keeping close to $f(t) = t$ for $t \leq n/2$, and then turning significantly smaller (see also Fig. 5).

Following the heuristic, we pick (1 2) as the first transposition and remove it, along with (2 3) and (2 4). Recalculating the weights, all vertices get the weight one and we can continue as we please. Note that while in this example, we could have chosen any vertex to begin with, in general we find acyclic subgraphs of different sizes if we choose the transpositions differently.

Although it seems that MAXIMAL APPROACH is not harder than TRANSPOSITION MEDIAN for $k = 3$, we have not been able to show this. However, we shall see in Section 6 that for the cases where computing the median is relevant, our approximation of MAXIMAL APPROACH usually attains the lower limit for the median distance, and hence solves TRANSPOSITION MEDIAN.

5. A closer look at expected distances

Not only is it easy to compute the transposition distance of a permutation, or between any two permutations for that matter, it is also fairly easy to compute the expected transposition distance of a product of t transpositions, taken randomly and independently from a uniform distribution on all transpositions. The following theorem gives the formula.

Theorem 7 (*Eriksen and Hultman [15]*). *Given a stochastic process $\{\pi^t\}_{t\geq 0}$, where $\pi^t \in \mathcal{P}_{nt}$, $\pi^t = \pi^{t-1}(a_t\ b_t)$, $(a_t\ b_t)$ is picked uniformly at random from $\mathcal{P}_{n1}$ and $\pi^0 = \mathrm{id}$, let $X_t = d_{\mathrm{trp}}(\pi^t)$. The expected transposition distance after t random transpositions in S_n is given by*

$$\mathbb{E}(X_t) = n - \sum_{k=1}^{n} \frac{1}{k} + \sum_{p=1}^{n-1} \sum_{q=1}^{\min(p,n-p)} a_{pq} \left(\frac{\binom{p}{2} + \binom{q-1}{2} - \binom{n-p-q+2}{2}}{\binom{n}{2}} \right)^t,$$

where

$$a_{pq} = (-1)^{n-p-q+1} \frac{(p-q+1)^2}{(n-q+1)^2(n-p)} \binom{n-p-1}{q-1} \binom{n}{p}.$$

The appearance of $\mathbb{E}(X_t)$ is given in Figs. 4 and 5. From the graphs, we draw the conclusion that for all n, $\mathbb{E}(X_t)$ is very close to t for $t \leq n/2$, whereafter the distance rapidly increases. We shall now use the same approach to obtain more results.

As mentioned in [15], computations similar to those that gave the expected value will also give the variance $\mathbb{V}(X_t)$, or equivalently the standard deviation $\mathbb{D}(X_t)$. An important difference is that the variance has a much more complicated expression and not needing it at that moment, the authors refrained from computing it.

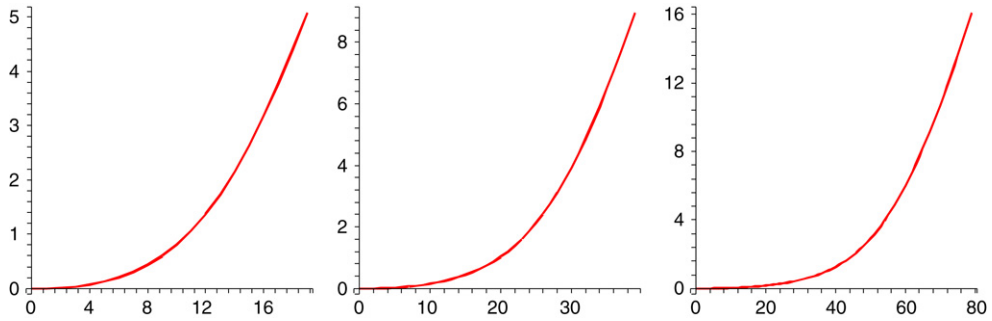


Fig. 5. The graphs show $t - \mathbb{E}(X_t)$ as a function of t, where X_t is as given in Theorem 7. From left to right, we have 20 genes, 40 genes and 80 genes. This difference is about 1 for $t = n/2$, and then increases rapidly.

We now find the standard deviation interesting enough to compute it. What we need to compute is

$$\begin{aligned}
\mathbb{V}(X_t) &= \mathbb{E}(X_t^2) - \mathbb{E}(X_t)^2 \\
&= \sum_{\lambda\vdash n} \chi^\lambda(1^n) \left(\frac{c(\lambda)}{\binom{n}{2}}\right)^t \sum_{\mu\vdash n} \frac{\chi^\lambda(\mu)(n-\ell(\mu))^2}{z_\mu} - \mathbb{E}(X_t)^2 \\
&= \sum_{\lambda\vdash n} \chi^\lambda(1^n) \left(\frac{c(\lambda)}{\binom{n}{2}}\right)^t \sum_{\mu\vdash n} \frac{\chi^\lambda(\mu)(n^2-2n\ell(\mu)+\ell(\mu)^2)}{z_\mu} - \mathbb{E}(X_t)^2 \\
&= n^2 + 2n(\mathbb{E}(X_t) - n) + \sum_{\lambda\vdash n} \chi^\lambda(1^n) \left(\frac{c(\lambda)}{\binom{n}{2}}\right)^t \sum_{\mu\vdash n} \frac{\chi^\lambda(\mu)\ell(\mu)^2}{z_\mu} - \mathbb{E}(X_t)^2 \\
&= n^2 + 2n(\mathbb{E}(X_t) - n) + (S - (\mathbb{E}(X_t) - n)) - \mathbb{E}(X_t)^2,
\end{aligned}$$

where

$$S = \sum_{\lambda\vdash n} \chi^\lambda(1^n) \left(\frac{c(\lambda)}{\binom{n}{2}}\right)^t \sum_{\mu\vdash n} \frac{\chi^\lambda(\mu)(\ell(\mu)^2 - \ell(\mu))}{z_\mu}.$$

The key to computing S lies in computing the inner sum over μ, the other factors being easier to handle. Eriksen and Hultman showed that

$$\sum_{\mu\vdash n} \frac{\chi^\lambda(\mu)(\ell(\mu)^2 - \ell(\mu))}{z_\mu} = \sum_{j=1}^{n-1}\sum_{k=1}^{n-j} \frac{1}{jk} \sum_T (-1)^{\mathrm{ht}(T)},$$

where the sum over T is over all border strip tableaux of shape $\lambda/(n-j-k)$ and type $(\max(j,k), \min(j,k))$. Border strip tableaux are formally defined in [28], but in essence a border strip tableau of this shape and type is a Ferrers diagram of $\lambda/(n-j-k)$ filled with $\max(j,k)$ ones and $\min(j,k)$ twos such that the squares filled with ones and twos, respectively, form connected skew shapes with no 2×2 subdiagrams, and such that removing all boxes numbered 2, we still have a valid Ferrers diagram. The boxes with ones thus form a ribbon along the border of $(n-j-k)$ and the boxes with twos form a ribbon along the border of all other boxes.

It follows that if $\lambda_4 > 2$, the sum is zero. For $\lambda_4 \leq 2$, there are usually many border strip tableaux to consider, which make computations complicated. There is probably no short formula for S, but a lengthy one is given in the Appendix.

The complexity of the formula for the standard deviation makes us content ourselves with showing its behaviour in a few graphs, collected in Fig. 6, where it is compared to $t - \mathbb{E}(X_t)$. We see that contrary to $t - \mathbb{E}(X_t)$, the standard deviation stays reasonably small for all t, showing that most $\pi \in \mathcal{P}_{nt}$ have $d_{\mathrm{trp}}(\pi)$ close to $\mathbb{E}(X_t)$.

The standard deviation takes its maximal value close to $t = n$. To see how fast this maximal value grows with n, we have plotted $\mathbb{D}(X_t)$ at $t = n$ for various n in Fig. 7. We find that the standard deviation grows much slower than the expectation, and that the growth pace seems to decrease as n increases. Since $\mathbb{D}(X_t)$ does not grow past 3

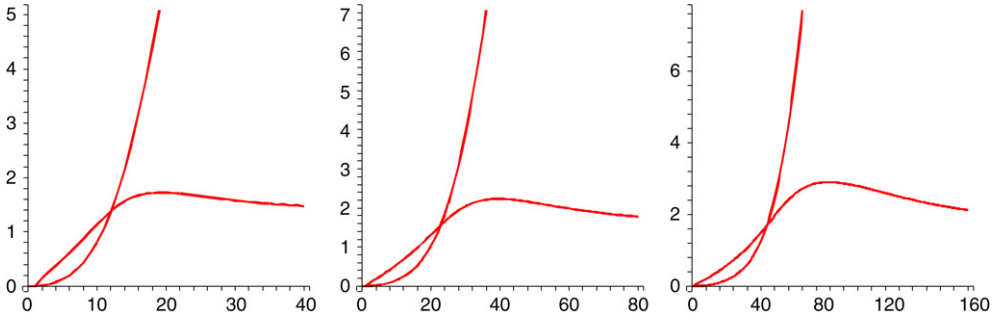


Fig. 6. The standard deviation $\mathbb{D}(X_t)$ of the transposition distance after t transpositions and $t-\mathbb{E}(X_t)$, with t on the abscissa. The standard deviation is the function that stays small for large t. From left to right, we have 20 genes, 40 genes and 80 genes. Note that the graph of $t-\mathbb{E}(X_t)$ gets steeper as n increases, in comparison to $\mathbb{D}(X_t)$. We see that the standard deviation is significantly smaller than $t-\mathbb{E}(X_t)$ for $t \geq 3n/4$.

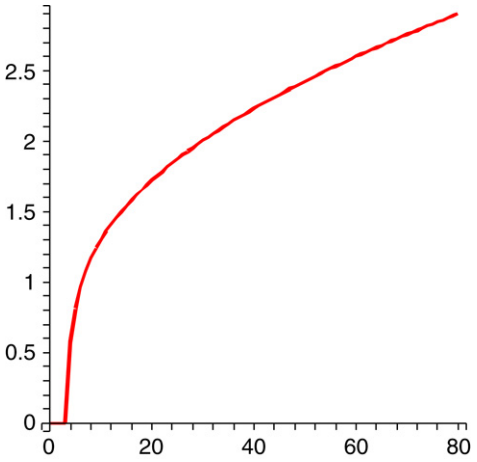


Fig. 7. The standard deviation $\mathbb{D}(X_t)$ at $t = n$ with n on the abscissa. We see that $\mathbb{D}(X_n)$ grows much slower than $\mathbb{E}(X_n)$, which is not too far from n.

even for $n = 80$, we draw the conclusion that even for permutations with many genes, the distance of products of t transpositions is well gathered about its expected value.

If t is small compared to n, both $t-\mathbb{E}(X_t)$ and $\mathbb{V}(X_t)$ are less than one, so it is very likely that $d(\pi) > t-6$ for $\pi \in \mathcal{P}_{nt}$. If we assume that $d(\pi) > t-6$ for all $\pi \in \mathcal{P}_{nt}$, then we can compute $\mathbb{P}_{X_t}(t) = \mathbb{P}(X_t = t)$. To be more precise, we use the expected value

$$\begin{aligned}\mathbb{E}(X_t) &= \sum_k k\mathbb{P}_{X_t}(k)\\ &= (t-4)\mathbb{P}_{X_t}(t-4) + (t-2)\mathbb{P}_{X_t}(t-2) + t\mathbb{P}_{X_t}(t)\\ &= t - 2\mathbb{P}_{X_t}(t-2) - 4\mathbb{P}_{X_t}(t-4)\end{aligned}$$

and the variance

$$\begin{aligned}\mathbb{V}(X_t) &= \sum_k k^2\mathbb{P}_{X_t}(k) - (\mathbb{E}(X_t))^2\\ &= t^2 + (4-4t)\mathbb{P}_{X_t}(t-2) + (16-8t)\mathbb{P}_{X_t}(t-4) - (\mathbb{E}(X_t))^2,\end{aligned}$$

to obtain, by linearity,

$$\mathbb{P}_{X_t}(t-4) = \frac{\mathbb{V}(X_t) + (\mathbb{E}(X_t) - t + 1)^2 - 1}{8},$$

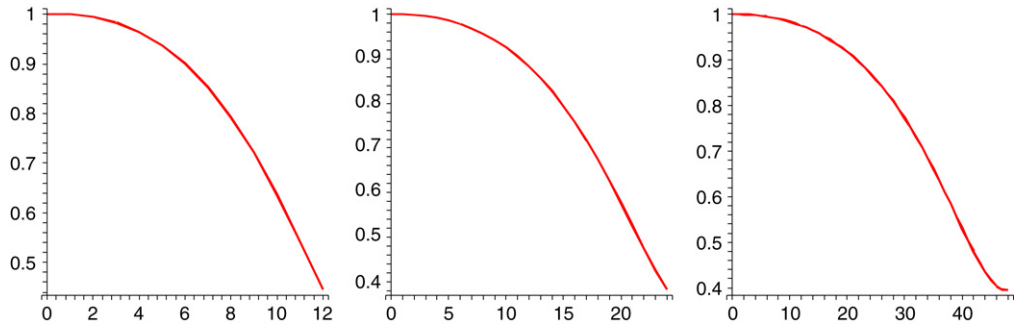


Fig. 8. The approximate probability that $d(\pi) = t$ for $\pi \in \mathcal{P}_{nt}$, with t on the abscissa. From left to right, we have 20 genes, 40 genes and 80 genes. The increasing derivative to the far right of the 80 gene graph is an artifact due to increasing approximation errors as t increases.

and

$$\mathbb{P}_{X_t}(t-2) = -\frac{\mathbb{V}(X_t) + (\mathbb{E}(X_t) - t + 2)^2 - 4}{4},$$

which gives

$$\mathbb{P}_{X_t}(t) = \frac{\mathbb{V}(X_t) + (\mathbb{E}(X_t) - t + 3)^2 - 1}{8}.$$

Our calculations have given us, with small error, the probability that all t transpositions have increased the distance to the identity. In Fig. 8, we see that for t up to about $n/4$, this probability is almost one. We then see a quite steep decrease to about $n/2$. For larger t, we do not expect this approximation to be valid, but reasonably, the probability should decrease to almost zero for $t = n - 1$.

In fact, this probability $\mathbb{P}_{X_{n-1}}(n-1)$ is computable. Using the same machinery that gave the expectation and standard deviation, it is possible to compute the probability that $X_t = n - j$ for any $1 \le j \le n$. What we need to compute is

$$\sum_{\ell(\mu)=j} \frac{\boldsymbol{e}_{1^n} M_n^t \boldsymbol{e}_\mu}{\binom{n}{2}^t} = \sum_{\lambda \vdash n} \chi^\lambda(1^n) \left(\frac{c(\lambda)}{\binom{n}{2}}\right)^t \sum_{\ell(\mu)=j} \frac{\chi^\lambda(\mu)}{z_\mu}.$$

Since $\chi^\lambda(1^n)$ and $c(\lambda)$ are known, we need only compute the rightmost sum.

Theorem 8. *Let $1_{n,j}$ denote the characteristic function of the set of integer partitions $\mu \vdash n$ such that $\ell(\mu) = j$. Then we have*

$$\sum_\mu \frac{\chi^\lambda(\mu) 1_{n,j}}{z_\mu} = \frac{1}{j!} \sum_{A_{n,j}} \frac{1}{k_1 k_2 \cdots k_j} \sum_T (-1)^{\mathrm{ht}(T)}.$$

In the first sum, $A_{n,j}$ denotes the set of all j-tuples $(k_1, k_2, \ldots, k_j)$ such that $k_i \ge 1$ and $\sum_i k_i = n$. The second sum is over all border strip tableaux T of shape λ and type $(k_1, k_2, \ldots, k_j)$.

Proof. Turning to symmetric functions, we need to find c_λ which fulfill the equation

$$\sum_{\lambda \vdash n} \frac{p_\lambda 1_{n,j}}{z_\lambda} = \sum_{\lambda \vdash n} c_\lambda s_\lambda.$$

Starting, as in [15], with Eq. (7.20) from [28], that is

$$\sum_\lambda \frac{p_\lambda(x) p_\lambda(y)}{z_\lambda} = \exp \sum_{n \ge 1} \frac{p_n(x) p_n(y)}{n},$$

we let the the first j y-variables be one and the rest be zero. We then differentiate j times with respect to t to obtain

$$\sum_{\lambda} \frac{\ell(\lambda)!}{(\ell(\lambda)-j)!} \frac{p_\lambda(x) t^{\ell(\lambda)-j}}{z_\lambda} = \left(\sum_{n\geq 1} \frac{p_n(x)}{n}\right)^j \exp \sum_{n\geq 1} t \frac{p_n(x)}{n}.$$

Putting $t = 0$ reduces the equation to

$$j! \sum_{\ell(\lambda)=j} \frac{p_\lambda(x)}{z_\lambda} = \left(\sum_{n\geq 1} \frac{p_n(x)}{n}\right)^j,$$

and considering terms of degree n only gives

$$j! \sum_{\lambda} \frac{p_\lambda(x) 1_{n,j}}{z_\lambda} = \sum_{A_{n,j}} \frac{p_{k_1}(x) p_{k_2}(x) \cdots p_{k_j}(x)}{k_1 k_2 \cdots k_j}.$$

Eqs. (7.74) and (7.75) of [28] now transform the power sum symmetric functions p_k into Schur functions s_λ, giving

$$j! \sum_{\lambda} \frac{p_\lambda 1_{n,j}}{z_\lambda} = \sum_{\lambda \vdash n} s_\lambda \sum_{A_{n,j}} \frac{1}{k_1 k_2 \cdots k_j} \sum_{T} (-1)^{\mathrm{ht}(T)},$$

which is the symmetric function equivalent of the theorem.

The double sums in this theorem get messy when we try to plug in numbers, especially for large j, but for $j = 1$ we can state this corollary.

Corollary 9. *The probability that $X_t = n - 1$ after t transpositions is given by*

$$\frac{1}{n} \sum_{s=0}^{n-1} (-1)^s \binom{n-1}{s} \left(\frac{\binom{n-s}{2} - \binom{s+1}{2}}{\binom{n}{2}}\right)^t.$$

Proof. If we plug in $j = 1$ into Theorem 8, we get

$$\sum_{\lambda=(n)} \frac{\chi^\lambda(\mu)}{z_\lambda} = \frac{1}{n} \sum_{T} (-1)^{\mathrm{ht}(T)} = \frac{(-1)^s}{n},$$

since there is only one border-strip tableau of shape $\lambda = (n-s, 1^s)$ and none if $\lambda_2 \geq 2$. For $\lambda = (n-s, 1^s)$, it also follows that

$$\chi^\lambda(1^n) = \frac{n!}{\prod_{c\in\lambda} h_c} = \binom{n-1}{s},$$

via the hook-length formula, and

$$\left(\frac{c(\lambda)}{\binom{n}{2}}\right)^t = \left(\frac{\binom{n-s}{2} - \binom{s+1}{2}}{\binom{n}{2}}\right)^t,$$

since

$$c(\lambda) = \sum_{j=1}^{l(\lambda)} \sum_{i=1}^{\lambda_j} (i-j).$$

In Fig. 9, we have plotted $\mathbb{P}_{X_t}(n-1)$ for those t that do not by parity make the probability zero. We see that it does not take too long to come close to the limit $2/n$. It is reasonable to assume that when the variation distance between the distribution after t transpositions and the uniform distribution, restricted to this subset of $\mathfrak{S}_n$, approaches zero, the same variation distance on the whole of $\mathfrak{S}_n$ should also do so. This seems in accordance with the results in [11], where Diaconis and Shahshahani show that a very similar stochastic process approaches the uniform distribution at a speed relative to $e^{t/n}$, starting from $t = n \log n/2$.

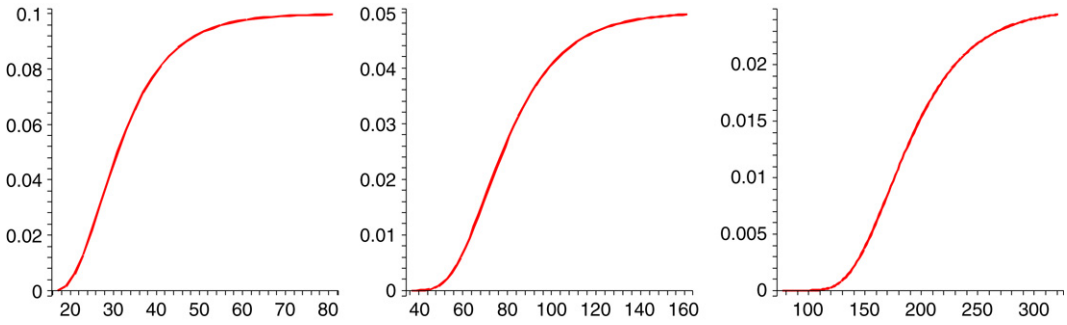


Fig. 9. The probability that $d(\pi) = n - 1$ for $\pi \in \mathcal{P}_{nt}$, with t on the abscissa. From left to right, we have 20 genes, 40 genes and 80 genes.

6. The reversal medians are different from the ancestor, and not unique

We are at last ready to make our point. Using the machinery introduced in the previous sections, we will now give convincing evidence that for moderately distant genomes, there are many medians, not necessarily close to each other. These medians will be of little aid in finding the ancestor of the genomes.

Assume that we have three permutations $\{\pi_1, \pi_2, \pi_3\}$, all in $\mathcal{P}_{40,10}$. Then, the pairwise differences, $\pi_1\pi_2^{-1}$ etc., all belong to $\mathcal{P}_{40,20}$. Looking at the probability that $d_{\text{trp}}(\pi) = t$ for $\pi \in \mathcal{P}_{nt}$, we can conclude that it is likely that $d(\pi_i) = 10$, for all i, but not $d(\pi_i, \pi_j) = 20$ for all i and j. This will prevent us from finding the evolutionary ancestor of $\{\pi_1, \pi_2, \pi_3\}$.

Say that we have $d(\pi_i) = 10$, and that $d(\pi_1, \pi_2) = d(\pi_1, \pi_3) = 18$ and $d(\pi_2, \pi_3) = 20$. Then, moving in from π_1, there are at most eight transpositions that reduce the distance to both π_2 and π_3. In fact, if we obtain the lower limit for the median distance, any median π has $d(\pi, \pi_1) = 8$ and thus $d(\pi) \geq 2$. If we do not obtain the lower limit, there may be medians that are closer to id, but the medians will be at some distance from each other and there is no known way to discriminate between them. Either way, the result is not satisfactory.

To see how bad the situation is, we have used the MAXIMAL APPROACH heuristic described above for computing medians of three permutations some distance apart. Examples of such computations are gathered in Table 1. The top group shows that if all three permutations π_i belong to $\mathcal{P}_{40,5}$, we usually obtain the sought ancestor. For permutations in $\mathcal{P}_{40,10}$, we see that while the ten transpositions move π_i away from id, they do not all move π_i away from the other two permutations. This introduces small errors in the results, errors that MAXIMAL APPROACH can not take the blame for.

For permutations in $\mathcal{P}_{40,15}$ the situation looks hopeless. In four of these five cases, our MAXIMAL APPROACH heuristic performs optimally, but the results are still very poor, with median candidates as far as 13 from the ancestor and 19 from each other. It is unlikely that any median computations on permutations this far apart can aid in the construction of a phylogenetic tree. Finally, we see that if the permutations are divided equally between these three sets, the behaviour is similar to the $\mathcal{P}_{40,10}$ case.

We have also conducted a more thorough search, making 2000 simulations for each category. The results are gathered in Table 2. In all four categories, the distance $d(\pi^i)$ from each permutation to the ancestor is fairly close to its maximal value. The maximal number of steps that MAXIMAL APPROACH can find has however gone significantly below its optimal value, especially for $\mathcal{P}_{40,15}$. It follows that even though MAXIMAL APPROACH does a great job, the median candidates are both far away from the ancestor and from each other, the $\mathcal{P}_{40,5}$ case excluded.

In the last row of the table we have added a new statistic, namely the maximal distance between median candidates from the same permutation π_i, run through the randomised version of MAXIMAL APPROACH. We see that for the $\mathcal{P}_{40,15}$ case, we can not even agree on a median candidate from each of the permutations.

Another approach to finding medians that scales terribly but has been surprisingly good for $n = 40$ is to improve on good median candidates by randomly picking genomes in their neighbourhood and introduce them as median candidates if they show comparable median distance. By iterating until no more progress seems to be made, we have found the data in Table 3. As predicted, for $\mathcal{P}_{40,5}$ id is often the sole median, but for $\mathcal{P}_{40,15}$, id is only a median in 1% of the simulations and distances between median candidates has been as large as 18 with an average at almost 10, showing the wide spread of medians. Perhaps even worse, among the medians there was one at the distance of

Table 1
Examples of median computations

$d(\pi^i)$			d^i_{opt}			$d^i_{\text{opt}} - d(\pi^i, \pi^i_{\text{opt}})$			$d(\pi^i_{\text{opt}})$			$d(\pi^i_{\text{opt}}, \pi^j_{\text{opt}})$		
5	5	5	5	5	5	0	0	0	0	0	0	0	0	0
5	5	5	5	5	5	0	0	0	0	0	0	0	0	0
5	5	5	4	6	4	0	0	0	1	1	1	2	2	0
5	5	5	5	5	5	0	0	0	0	0	0	0	0	0
5	5	5	5	5	5	0	0	0	0	0	0	0	0	0
10	10	10	11	9	9	0	0	0	1	1	1	2	2	2
10	10	10	9	9	11	0	0	0	1	1	1	2	2	2
10	10	10	10	10	10	0	0	0	0	0	0	0	0	0
10	10	10	9	9	11	0	0	0	1	1	1	2	2	2
10	8	10	9	9	9	0	0	0	1	1	1	2	2	2
15	15	13	14	14	14	0	0	0	5	3	3	8	8	6
15	15	15	13	11	13	0	0	0	8	8	6	12	12	10
15	15	15	13	13	15	0	0	1	6	8	13	10	17	19
15	15	15	15	13	15	0	0	0	6	6	8	12	10	12
15	15	15	14	12	14	2	2	2	7	9	7	14	14	14
5	10	15	6	9	14	1	2	2	2	3	3	5	5	6
5	10	15	6	9	14	0	0	0	1	1	1	2	2	2
5	10	15	5	10	15	0	0	0	2	0	0	2	2	0
5	8	15	5	8	15	0	0	0	0	0	0	0	0	0
5	10	15	6	7	14	0	0	0	3	3	3	4	6	6

We have provided five examples each from four different setups: the three permutations π_1, π_2 and π_3 belong to $\mathcal{P}_{40,5}$, $\mathcal{P}_{40,10}$, $\mathcal{P}_{40,15}$ and finally one each, respectively. The following output data are provided: the distance from π^i to the ancestor id, the distance from π^i to a potential median at the lower limit (d^i_{opt}), the number of steps short of d^i_{opt} our MAXIMAL APPROACH-algorithm stops (at the permutation π^i_{opt}), the distance between id and π^i_{opt}, and finally the difference between the results from our MAXIMAL APPROACH-algorithm on the three permutations.

Table 2
Averages of the indicated values, computed over 2000 simulations

	$\mathcal{P}_{40,5}$	$\mathcal{P}_{40,10}$	$\mathcal{P}_{40,15}$	Mix
$d(\pi^i)$	4.97	9.83	14.56	9.80
d^i_{opt}	4.92	9.47	13.03	9.41
$d^i_{\text{opt}} - d(\pi^i, \pi^i_{\text{opt}})$	0.00	0.12	0.68	0.17
$d(\pi^i_{\text{opt}})$	0.19	1.81	7.95	1.81
$d(\pi^i_{\text{opt}}, \pi^j_{\text{opt}})$	0.28	3.16	13.13	3.16
$\max_{j,m} d(\pi^i_{\text{opt}_j}, \pi^i_{\text{opt}_m})$	0.07	1.34	7.76	1.15

The categories are the same as in Table 1, except for the last one. In each simulation, we have, for each π^i, computed 30 approximations of the median by using the randomised version of our MAXIMAL APPROACH-algorithm. The value in the last row is the mean of the maximal pairwise difference between these. A high value indicates that the median candidates obtained from π^i are highly scattered.

Table 3
Averages of worst cases in each simulation, computed over 2000 simulations

	$\mathcal{P}_{40,5}$	$\mathcal{P}_{40,10}$	$\mathcal{P}_{40,15}$	Mix
$\sum d(\pi_{\text{med}}, \pi^i)$	14.78	28.56	39.69	28.37
$\min \sum d(\pi_{\text{med}}, \pi^i)$	11	22	31	22
$d(\pi_{\text{med}})$	0.07	1.26	5.32	1.12
$\max d(\pi_{\text{med}})$	2	8	14	10
$d(\pi^i_{\text{med}}, \pi^j_{\text{med}})$	0.21	2.21	9.46	2.09
$\max d(\pi^i_{\text{med}}, \pi^j_{\text{med}})$	3	10	18	10
$\mathbb{P}(\text{id} = \pi_{\text{med}})$	0.87	0.34	0.01	0.32

To exemplify, the third column shows that for $\mathcal{P}_{40,15}$, we have on average a median distance of 39.7, the smallest median distance being 31. On average, the largest distance to id among the medians is 5.3, peaking at 14, that is almost the evolutionary distance from id to the genomes in the test set. Again on average, the largest distance among medians found is 9.5, with a maximum at 18, and finally the proportion of simulations where the true ancestor id is at least one of the medians is 1%.

Table 4
Human, sea urchin and fruit fly mtDNA gene orders [25], together with two candidate medians with median distance 39

Human	26 13 17 12 -24 15 18 32 -2 -16 -3 -33 4 -28 7 5 1 10 19 25 22 11 29 14 20 -21 -8 6 30 -23 9 27 31
Sea urchin	26 4 25 22 5 1 -28 19 11 29 20 -21 6 9 27 8 30 23 -24 16 14 -2 32 3 -31 15 -7 33 10 13 17 12 18
Fruit fly	-26 -31 -27 12 -24 15 18 32 -3 -33 4 13 5 7 1 10 19 2 25 16 29 8 -9 -20 -11 -22 30 -23 21 6 28 -17 -14
Med cand 1	26 13 17 12 -24 15 18 32 -28 7 -6 21 -20 -14 -4 33 3 -8 16 5 1 10 19 25 22 11 29 -2 30 -23 9 27 31
Med cand 2	26 13 17 12 -24 15 18 32 25 22 11 29 14 -2 -19 -10 -1 -5 -3 -33 4 -28 7 16 20 -21 23 -30 -6 8 9 27 31

Their mutual distance is 11, which is very high in relation to their distances to the human gene order, which are 6 and 5, respectively.

14 from the true ancestor id. This is by and large the distance to the test set, showing that medians can be very peripheral.

7. Applications to real data

A good three genome data set is given by Sankoff et al. [25], consisting of human, sea urchin and fruit fly mtDNA (Table 4). Having only 33 genes, the latter two are at the great distance of 32 reversals from each other, but the distances to the human mtDNA is smaller (19 and 24, respectively). Grappa [22] has reported a median candidate with median distance of 43 reversals, whereas both the original authors and MGR [7] reports candidates with median distance 39. Since a lower limit of the median distance is 38, these are good candidates, but can we rely on them for computing phylogenetic trees?

Sankoff et al. found 3 medians and MGR contributed with another one. By looking at genomes close to these, we have found some 10 more. Already this indicates that the genome is far from unique, but these (candidate) medians may still be very close. On the contrary, two of them, one from the original article and one new, are separated by no less than 11 reversals. This could be compared for instance with their distance to the human gene order, which is 6 and 5 reversals, respectively. We strongly suspect that if a median of the three mitochondrial gene orders was to be used for phylogeny reconstruction, the result would depend heavily on which median were used. After all, our simulations indicate strongly the possibility that the ancestor is not a median. If we were to pick the median furthest away from the ancestor, this could be as far from the ancestor as the human mtDNA is from sea urchin and fruit fly, which introduces more problems than it solves.

Appendix. The sum in the variance

As described, we need only consider partitions λ for which $\lambda_4 \leq 2$. To keep track of the heights of the border strips, we have divided the tableau into five parts as in Fig. A.1, that is $p = \lambda_1$, $q = \lambda_2 - 1$, $r = \lambda_3 - 2$, $s = \lambda'_1 - 1$ and $t = \lambda'_2 - 2$.

Now, computing S is just a matter of good book-keeping. We start with the simplest partitions, having $q = r = u = 0$, and gradually add more complexity. In this formula, we have used Iverson's convention as expressed in [17], that is [logical statement] equals one if the logical statement is true and zero otherwise.

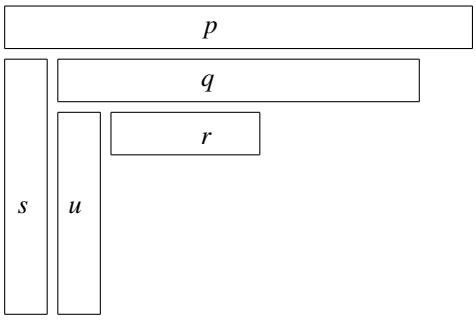


Fig. A.1. Partitioning the tableau into five parts.

$$
\begin{aligned}
S = {} & \sum_{j=1}^{n-1}\sum_{k=1}^{n-j}\frac{1}{jk} + \sum_{s=1}^{n-1}[p+s=n]\frac{n!(-1)^s}{(p-1)!s!(p+s)}\left(\frac{\binom{p}{2}-\binom{s+1}{2}}{\binom{n}{2}}\right)^t \\
& \times\left(\sum_{k=1}^{s-1}\frac{1}{(s-k)k} - \frac{2}{s}\sum_{k=1}^{p-1}\frac{1}{p-k} - \sum_{k=0}^{s-1}\frac{1}{(p+k)(s-k)}\left([s-p-2k\le 0]+[s-p-2k<0]\right)\right. \\
& \quad \left. + \sum_{k=1}^{p-1}\frac{1}{(p-k)(s+k)}\left([p-s-2k\le 0]+[p-s-2k<0]\right)\right) \\
& + \sum_{s=1}^{n-1}\sum_{q=1}^{\lfloor\frac{n-s-1}{2}\rfloor}[p+q+s=n]\frac{n!(p-q)(-1)^s}{p!q!(s-1)!(p+s)(q+s)}\left(\frac{\binom{p}{2}+\binom{q}{2}-\binom{s+1}{2}}{\binom{n}{2}}\right)^t \\
& \times\left(\frac{2}{p+s}\sum_{k=1}^{q-1}\frac{1}{q-k} - \frac{2}{q+s}\sum_{k=1}^{p-1}\frac{1}{p-k}\right. \\
& \quad + \sum_{k=0}^{p-1}\frac{1}{(p-k)(s+k)}\left([p-s-2k\le 0]+[p-s-2k<0]\right) \\
& \quad - \sum_{k=0}^{q-1}\frac{1}{(q-k)(s+k)}\left([q-s-2k\le 0]+[q-s-2k<0]\right) \\
& \quad - \sum_{k=1}^{s-1}\frac{1}{(p+k)(s-k)}\left([s-p-2k\le 0]+[s-p-2k<0]\right) \\
& \quad \left. + \sum_{k=1}^{s-1}\frac{1}{(q+k)(s-k)}\left([s-q-2k\le 0]+[s-q-2k<0]\right)\right) \\
& + \sum_{u=0}^{\lfloor\frac{n-4}{2}\rfloor}\sum_{s=u+1}^{n-u-3}\sum_{q=1}^{\lfloor\frac{n-s-u-1}{2}\rfloor}[p+q+s+u=n]\frac{n!(p-q)(s-u)(-1)^{s+u}}{(p-1)!(q-1)!s!u!(p+s)(q+s)(p+u)(q+u)} \\
& \times\left(\frac{\binom{p}{2}+\binom{q}{2}-\binom{s+1}{2}-\binom{u+1}{2}}{\binom{n}{2}}\right)^t\left(\frac{2}{(p+s)(q+u)} - \frac{2}{(p+u)(q+s)}\right) \\
& + \sum_{r=0}^{\lfloor\frac{n-6}{3}\rfloor}\sum_{q=r+1}^{\lfloor\frac{n-r-4}{2}\rfloor}\sum_{p=q+1}^{n-r-q-3}\sum_{u=1}^{\lfloor\frac{n-r-q-p-1}{2}\rfloor}[p+q+r+s+u=n] \\
& \times\frac{n!(p-q)(p-r)(q-r)(s-u)(-1)^{s+u}}{p!q!r!(s-1)!(u-1)!(p+s)(q+s)(r+s)(p+u)(q+u)(r+u)} \\
& \times\left(\frac{\binom{p}{2}+\binom{q}{2}+\binom{r}{2}-\binom{s+1}{2}-\binom{u+1}{2}}{\binom{n}{2}}\right)^t \\
& \times\left(\frac{2}{(q+s)(r+u)} - \frac{2}{(p+s)(r+u)} + \frac{2}{(r+s)(p+u)} - \frac{2}{(r+s)(q+u)}\right. \\
& \left. + [r>0]\left(\frac{2}{(p+s)(q+u)} - \frac{2}{(q+s)(p+u)}\right)\right).
\end{aligned}
$$

References

[1] D. Bader, B. Moret, M. Yan, A linear-time algorithm for computing inversion distance between signed permutations with an experimental study, Journal of Computational Biology 8 (5) (2001) 483–491.

[2] V. Bafna, P. Pevzner, Genome rearrangements and sorting by reversals, SIAM Journal of Computing 25 (1996) 272–289.

[3] V. Bafna, P. Pevzner, Sorting by transpositions, SIAM Journal of Discrete Mathematics 11 (1998) 224–240.
[4] A. Bergeron, A very elementary presentation of the Hannenhalli–Pevzner theory, in: Proceedings of the 12th Annual Symposium on Combinatorial Pattern Matching, in: LNCS, vol. 2089, Springer-Verlag, 2001, pp. 106–117.
[5] P. Berman, T. Fujito, Approximating independent sets in degree 3 graphs, in: Proc. 4th Workshop on Algorithms and Data Structures, in: LNCS, vol. 955, Springer-Verlag, 1995, pp. 449–460.
[6] M. Blanchette, T. Kunisawa, D. Sankoff, Parametric genome rearrangement, Gene 172 (GC) (1996) 11–17.
[7] G. Bourque, P. Pevzner, Genome-scale evolution: Reconstructing gene orders in the ancestral species, Genome Research 12 (2002) 26–36.
[8] A. Caprara, Formulations and hardness of multiple sorting by reversals, in: Proceedings of the Third Annual International Conference on Computational Molecular Biology, RECOMB 99, ACM Press, 1999, pp. 84–93.
[9] A. Caprara, On the practical solution of the reversal median problem, in: Algorithms in Bioinformatics, Proceedings of WABI 2001, in: LNCS, vol. 2149, Springer-Verlag, 2001, pp. 238–251.
[10] D. Christie, Genome rearrangement problems, Ph.D. Thesis, Department of Computer Science, University of Glasgow, 1998.
[11] P. Diaconis, M. Shahshahani, Generating a random permutation with random transpositions, Zeitschrift für Wahrscheinlichkeitstheorie und verwandte Gebiete 57 (1981) 159–179.
[12] I. Elias, T. Hartman, A 1.375-approximation algorithm for sorting by transpositions, in: Proceedings of WABI 2005, in: LNCS, vol. 3692, Springer-Verlag, 2005, pp. 204–214.
[13] N. Eriksen, Approximating the expected number of inversions given the number of breakpoints, in: Algorithms in Bioinformatics, Proceedings of WABI 2002, in: LNCS, vol. 2452, Springer-Verlag, 2002, pp. 316–330.
[14] N. Eriksen, $(1+\varepsilon)$-approximation of sorting by reversals and transpositions, Journal of Theoretical Computer Science 289 (2002) 517–529.
[15] N. Eriksen, A. Hultman, Estimating the expected reversal distance after a fixed number of reversals, Advances of Applied Mathematics 32 (2004) 439–453.
[16] N. Eriksen, Combinatorial problems in comparative genomics, Ph.D. Thesis, Department of Mathematics, Royal Institute of Technology, Stockholm, 2003.
[17] R. Graham, D. Knuth, O. Patashnik, Concrete Mathematics: A Foundation for Computer Science, Addison-Wesley, Boston, 1994.
[18] S. Hannenhalli, P. Pevzner, Transforming cabbage into turnip (polynomial algorithm for sorting signed permutations with reversals), Journal of the ACM 46 (1999) 1–27.
[19] B. Moret, D. Bader, S. Wyman, T. Warnow, M. Yan, A new implementation and detailed study of breakpoint analysis, in: Proceedings of the Pacific Symposium of Biocomputing, PSB 01, World Scientific, 2001, pp. 583–594.
[20] B. Moret, A. Siepel, J. Tang, T. Liu, Inversion medians outperform breakpoint medians in phylogeny reconstruction from gene-order data, in: Algorithms in Bioinformatics, Proceedings of WABI 2002, in: LNCS, vol. 2452, Springer-Verlag, 2002, pp. 521–536.
[21] B. Moret, J. Tang, L. Wang, T. Warnow, Steps toward accurate reconstructions of phylogenies from gene-order data, in: Computational Biology, Journal of Computer and Systems Sciences 63 (2002) 508–525 (special issue).
[22] B. Moret et al., Grappa (Genome Rearrangements Analysis under Parsimony and other Phylogenetic Algorithms), version 2.0. Available for free download at: http://www.cs.unm.edu/~moret/GRAPPA/.
[23] I. Pe'er, R. Shamir, The median problems for breakpoints are **NP**-complete, in: Electronic Colloquium on Computational Complexity, TR98-071, 1998.
[24] D. Sankoff, M. Blanchette, Multiple genome rearrangement and breakpoint analysis, Journal of Computational Biology 5 (1998) 555–570.
[25] D. Sankoff, G. Sundaram, J. Kececioglu, Sterner points in the space of geneome rearrangements, International Journal of the Foundations of Computer Science 7 (1996) 1–9.
[26] A. Siepel, An algorithm to find all sorting reversals, in: Proceedings of the Sixth Annual International Conference on Computational Molecular Biology, RECOMB 02, ACM Press, 2002, pp. 281–290.
[27] A. Siepel, B. Moret, Finding an optimal inversion median: Experimental results, in: Algorithms in Bioinformatics, Proceedings of WABI 2001, in: LNCS, vol. 2149, Springer-Verlag, 2001, pp. 189–203.
[28] R.P. Stanley, Enumerative Combinatorics, vol. 2, Cambridge University Press, New York, Cambridge, 1999.
[29] L.-S. Wang, T. Warnow, Distance based genome rearrangement phylogeny, in: O. Gascuel (Ed.), Mathematics of Evolution and Phylogeny, Oxford University Press, New York, 2005, pp. 353–383.